On Visual Detection of Highly-occluded Objects for Harvesting Automation in Horticulture

Inkyu Sa*, Christopher McCool*, Christopher Lehnert*, and Tristan Perez*

***Abstract*— Developing accurate and reliable crop detection algorithms is an important step for harvesting automation in horticulture. This paper presents a novel approach to visual detection of highly-occluded fruits. We use a conditional random field (CRF) on multi-spectral image data (colour and Near-Infrared Reflectance, NIR) to model two classes: crop and background. To describe these two classes, we explore a range of visual-texture features including local binary pattern, histogram of oriented gradients, and learn auto-encoder features. The proposed methods are evaluated using hand-labelled images from a dataset captured on a commercial capsicum farm. Experimental results are presented, and performance is evaluated in terms of the Area Under the Curve (AUC) of the precision-recall curves. Our current results achieve a maximum performance of 0.81 AUC when combining all of the texture features in conjunction with colour information.**

I. INTRODUCTION

The United Nations (UN) predicts that food production may have to double by 2050 in order to cope with the expected population growth [1], [2]. This increased food production needs to occur despite limited supply of new arable land as well as growing difficulties in sourcing skilled farm labour. To meet these challenges, farm productivity must increase dramatically. Automating agricultural processes such as planting, harvesting, weeding and inspection using robotics will play a key role in improving farm productivity by increasing crop quality and reducing input costs. Recent research in robotics has made considerable progress towards the goal of developing viable broadacre [3] and horticultural robots [4].

In this work, we are interested in the task of automatic harvesting using vision technology, specifically accurate visual detection of fruit. Developing an accurate fruit detection system is an important first step in developing automated fruit harvesting robots since as this is the front-end perception system prior to subsequent manipulation and grasping systems—if a fruit is not detected or seen it cannot be picked.

We present preliminary results for a novel approach to visually detecting capsicums which makes use of both colour and texture features within a Conditional Random Field (CRF) framework. In particular, we explore the use of several texture features including local binary pattern, histogram of oriented gradients, and learn auto-encoder features. We evaluate these features on imagery gathered from a commercial farm. These images are cluttered, dynamic and contain many highly-occluded fruit (capsicum) as shown in Fig. 1. This is in contrast to similar previous work [4] where the images were acquired in a controlled glasshouse environment with a static background and favourable lighting.

Fig. 1. Colour and NIR image of capsicums (pepers). These figures show an instance of experiment scene that is complex and reasonably cluttered. Some capsicums are nicely located at the centre whereas others are highly-occluded by leaves and capsicums.

The contributions of this paper are as follows:

- Analysis of a range of texture features on capsicum detection, specifically Histogram of Oriented Gradients (HOG), Sparse Auto Encoder (SAE), and Local Binary Pattern (LBP).
- Comparison and evaluation of the proposed approach through AUC and F1 score on a challenging real-world dataset.

The remainder of the paper is structured as indicated in the following. Section II introduces related work and background. Section III describes image features that are utilised for capsicum detection. We present our experimental results and discussion in Section IV. Conclusions are drawn in Section V.

II. RELATED WORK/BACKGROUND

In this section, we present a review of literature on crop detection and classification. As previously mentioned, several research groups have shown crop detection using vision. Yamamoto et al. [5] developed a tomato detection system using prior information such as the colour, shape, texture and size of tomatoes in images. Decision-tree based pixel segmentation and random forest blob segmentation were performed. An overexposed image region caused by camera flash was detected to determine an individual tomato from a multi-tomato blob. Bac et al. [4] attempted to construct an obstacle map of capsicums plants for manipulator planning using multispectral imagery. They used a classifier and regression tree (CART) classifier to learn how to segment each pixel into soft and hard parts of the plants using a range of multi-spectral features and an entropy based texture feature. Unfortunately, the results were insufficient for manipulator planning and grasping. Nuske et al. [6] demonstrated grape

* Science and Engineering Faculty, Queensland University of Technology, Brisbane, Australia. i.sa@qut.edu.au, c.mccool@qut.edu.au, c.lehnert@qut.edu.au, tristan.perez@qut.edu.au

detection and yield estimation using a Radial Symmetry Transform which is often used for biometric iris identification. These authors also investigated the use of multiple flashes to estimate depth that can provide more accurate edge extraction and yield better grape detection. Wang et al. [7] utilised colour and specular reflection visual cues for apple detection and yield estimation. They estimated the 3D locations of apples in order to avoid over-counting the fruits appearing in multiple images and on the opposite side of a tree. Recently, Hung et al. [8] presented almond detection using feature learning and a CRF framework with colour and IR images and reported promising results for almond segmentation. This work is closely related to our method presented in this paper, however, we explore a range of other texture features and apply this to do a different target crop.

III. Methodology

We explore three features related to texture: a Histogram of Oriented Gradients (HOG), a learnt Sparse Auto Encoder (SAE) feature, and Local Binary Pattern (LBP) in order to investigate the impact of each feature for capsicum detection. In addition, overview of back-end supervised segmentation algorithm is presented.

A. Histogram of Oriented Gradients (HOG) feature

HOG are feature descriptors that have been widely used for object detection [9]. The HOG feature describes the distribution of local gradient magnitudes and their orientations. This is achieved by first dividing an image into small cells (e.g., 8×8 pixels). For each cell, a histogram of edge orientations is calculated for the individual pixels in each cell. Next, contrast-normalisation is performed in order to cope with illumination changes. Finally, blocks consisting of 2×2 cells are defined and a local histogram is accumulated over the block. These blocks overlap by 50%.

B. Sparse Auto Encoder (SAE) feature

A Sparse Auto Encoder [10] is an unsupervised feature learning approach based on neural networks. The objective in training a neural network is to optimise a set of unknown parameters such as weights, $\mathbf{W}$ and bias, $\mathbf{b}$ given input, $\mathbf{x}$ and training data, $\mathbf{y}$. There are two main steps to achieve this goal: feedforward and backpropagation. The first step can be done by propagating the sum of output values from neurons to the next cascaded layer as

$$\begin{aligned}\mathbf{z}^{(l+1)} =& \mathbf{W}^{(l)}\mathbf{a}^{(l)} + \mathbf{b}^{(l)} \\ h_{\mathbf{W},\mathbf{b}}(\mathbf{x}) =& \mathbf{a}^{(l+1)} = f(\mathbf{z}^{(l+1)})\end{aligned}$$

where $\mathbf{a}^{(l)}$ is activation of in layer l=[1, 2, 3] of the network, thus we can say $\mathbf{a}^{(1)} = \mathbf{x}$. The function $f(\cdot)$ is the sigmoid function with output range $[-1, 1]$. $h_{\mathbf{W},\mathbf{b}}(\mathbf{x})$ is the final activation with function of $\mathbf{W}$ and $\mathbf{b}$ given $\mathbf{x}$. The second backpropagation step iteratively optimises the following cost function, given m training samples,

$$\begin{aligned}\mathbf{J}(\mathbf{W},\mathbf{b}) = &\left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\|h_{\mathbf{W},\mathbf{b}}(\mathbf{x}^{(i)}) - \mathbf{y}^{(i)}\|^2\right)\right] \\ &+ \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(\mathbf{W}_{ij}^{(l)}\right)^2,\end{aligned} \quad (1)$$

where i and j indicate an index of a neuron in layer $l+1$ and l respectively. n_l is number of layers in the network and s_l is number of neurons in layer l.

The first term of Eq. (1) denotes an average of sum of squares error and the second is a weight decay term controlled by the parameter λ. The Auto Encoder imposes a constraint of $\mathbf{x} = \mathbf{y}$ which implies the learnt parameters $\mathbf{W}$ and $\mathbf{b}$ are optimised to make the input and output identical. This constraint transforms a supervised neural networks into an unsupervised network. In addition, if we introduce another constraint that $\mathbf{a}^{(l)} \approx -1$ that discourages activation of neurons. As a result, the unknown parameters are optimised so as to be inactive whenever possible. Our Sparse Auto Encoder takes as input an 8×8, window which results in 64 pixel intensities, and we have a single hidden layer consisting of 25 neurons.

C. Local Binary Pattern (LBP) feature

The Local Binary Pattern is a simple and powerful feature descriptor [11] [12] that is able to describe texture features by simply calculating a binary pattern. This binary pattern is computed by comparing a pixel to its neighbouring pixels. There are two control parameters: P denotes the number of sampling points (referred to as neighbourhood pixels before). R is the radius or between the centre point and sampling points. We can model these sampling points ($\mathrm{u_p}$,$\mathrm{v_p}$) centred at (u,v) as

$$\mathrm{u_p} = \mathrm{u} + R\cos(\frac{2\pi p}{P}), \quad \mathrm{v_p} = \mathrm{v} - R\sin(\frac{2\pi p}{P})$$

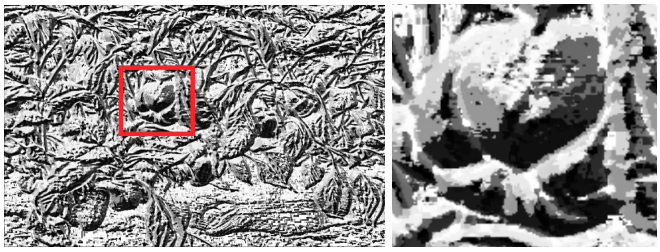
where the scalar p is an integer taking values in [0,P-1]. In case that a sampling point is not centred on a pixel, the resulting pixel values is calculated with bilinear interpolated. The binary pattern for the centre pixel (u,v) of image $I(\mathrm{u},\mathrm{v})$ can be computed as

$$LBP_{\mathrm{P,R}}(\mathrm{u},\mathrm{v}) = \sum_{p=0}^{P-1} s(I(\mathrm{u},\mathrm{v}) - I(\mathrm{u_p},\mathrm{v_p}))2^p$$

where s(·) is the thresholding function as defined

$$s(z) = \begin{cases} 1 & \text{if } z \geq 0 \\ 0 & \text{if } z < 0 \end{cases}.$$

Fig. 2 shows an LBP encoded image with parameters (P=4, R=1) and a closer look at an image region containing a capsicum. Qualitatively speaking, this image appears to show a distinguishable texture near capsicums and looks to be promising candidate for capsicum detection.



(a) (b)

Fig. 2. (a) is a LBP encoded image containing a capsicum (red box) and (b) is its zoomed in view.

D. Conditional Random Fields (CRFs) framework

It is common to represent an image using graphical models such as a either Hidden Markov Models (HMM), Maximum Entropy Markov Model (MEMM), or Conditional Random Fields (CRF). We choose to use a CRF [13], which takes into consideration neighbouring labels and pixels [14]. Below we provide a brief overview overview of CRFs, more details can be found in [15] (chapter 5.2.1).

The graphical model of an image may be written as $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ where $\mathbf{V}$ is all of the pixel locations such that the i-th pixel, x_i, corresponds to the vertices of the graph and $\mathbf{E}$ is an edge vector representing the relationship of adjacent pixels. The objective of a CRF is to estimate the label $\ell_i \in \mathbf{L}$ for every pixel location $i \in \mathbf{V}$. ℓ_i denotes the particular object class $\in [1...k]$ and in our case the number of classes is $k = 2$ (i.e., background and capsicum). Image segmentation can be achieved by minimising the energy-like function [8] of the graph given by

$$\begin{aligned} E(\ell) =& E_{\text{u}}(\ell) + E_{\text{p}}(\ell_i, \ell_j, x_i, x_j) - log(Z(x)) \\ =& \sum_{i \in \mathbf{V}} \sum_{f} w_{\text{f}} \psi_{\text{f}}(x_i|\theta_{\text{f}}) + \sum_{i,j \in \mathbf{E}} w_{i,j}(x_i, x_j)(\ell_i - \ell_j) \\ & - log(Z(x)). \end{aligned} \tag{2}$$

The first term of Eq. (2) represents the unary potential that is the likelihood of a pixel x_i having a label ℓ_i given feature θ_{f}. The second term is the pairwise potential that measures the coherence of the neighbouring pixel labels. The last term is the partition function.

Features described in the previous sections are fed into this CRF that is trained in a supervised manner using a hand-labelled dataset. Details about this dataset are given in the next section.

IV. Experimental Results

A. Experimental setup

In this section the experimental setup, as well as training and evaluation steps, are presented. An industrial camera, AD-130GE manufactured by JAI, is utilised for capsicum data collection. This prism-based two $1/3''$ CCD multi-spectral camera can record registered colour and NIR imagery simultaneously with a resolution of 1296×964; the colour response range is $380 - 700\,\text{nm}$ and NIR response range is $700 - 1000\,\text{nm}$. This camera is mounted on the QUT Video cart for data capture shown in Fig 3.

Fig. 3. QUT Video cart used for capsicum data collection [17].

Using one sequence of capsicum images we hand-labelled 20 pairs of RGB and NIR images and evaluated our approaches on this data. We randomly select 10 image pairs to performing training (of the CRF and Sparse Auto Encoder) and the remaining 10 image pairs were used to evaluate the performance of our capsicum detector testing.

For the classifier evaluation, we adopt two measures: AUC of precision-recall curve and the harmonic F1 score that measures the accuracy of capsicum detector. The latter can be simply calculated by $\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$ [16] . Precision and recall are first computed using the hand-labelled images and the output of the trained CRF (likelihood map or marginal image). To calculate the harmonic F1 score, we then choose the threshold τ where the precision and recall are equal. This threshold (τ) is then applied to the likelihood map in order to obtain the prediction image, as shown in Fig. 6.

B. Results with colour and NIR imagery

From our results in Fig. 4, it can be seen that the LBP features provide considerably better detections than either the HOG or SAE features. We believe that the poor performance of the HOG feature is due to the fact that this feature was designed for detection of structured objects (e.g., pedestrians, horses, bicycles, and motor bikes), whereas, capsicums do not have a high degree of distinguishable structure. The poor performance of the SAE feature is attributed to the limited amount of training data and we believe that supplying even more training data will yield improved results.

Finally, we combine all aforementioned features along with colour using the hue-saturation-value (HSV) representation to obtain our best performing system. With this approach we obtained a considerable improvement achieving and AUC of 0.812 compared to 0.730 when using just the LBP feature. Using this system we present an example output in Fig. 6. Finally, to explore the consistency of our result we apply the threshold (τ) and plot the F1 score for the 10 testing images as shown in Fig. 5. It can be seen that on average we obtain good performance, however, the variance is quite high meaning that we have good detections for some images and poor detections for other images. Future work will examine how to improve the system to obtain more consistent detection results.

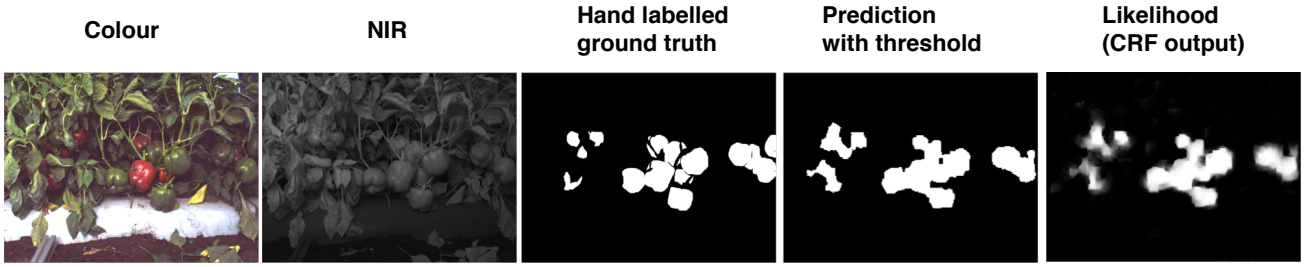


Fig. 6. An instance of capsicum segmentation results. Each column is an image corresponding to the label on top.

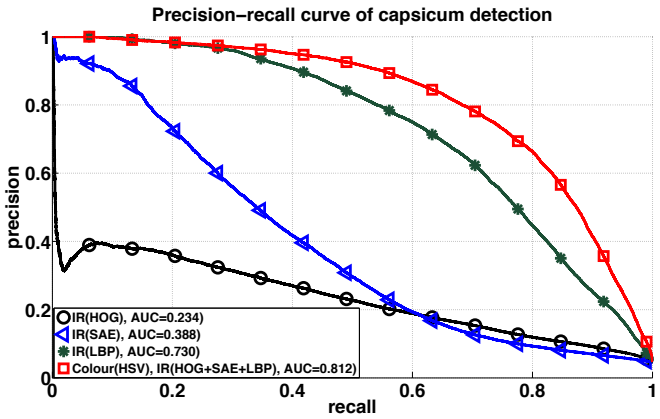


Fig. 4. Precision-recall curve for all possible thresholds. The threshold where precision and recall are identical is chosen to generate prediction image. Red line indicates the best performance with colour and NIR features.

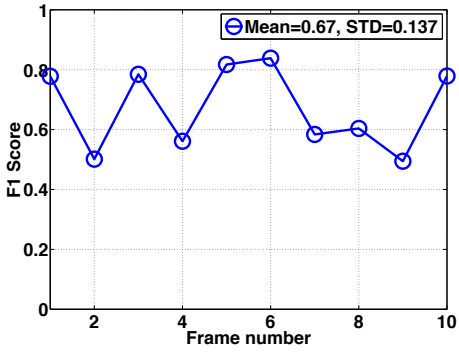


Fig. 5. F1 score for each frame and the mean and standard deviation.

V. Conclusions and Future Works

In this paper, we examine the impact of using different texture features for capsicum detection. We show experimental results for capsicum that highlight that the HOG and SAE features provide considerably worse performance than using the LBP feature. We also find that incorporating colour information and combining the three texture features yields the best performance with respect to an AUC, which attained of 0.812 for the example considered.

Future work will concentrate on two particular issues. First, more data will be collected and annotated so that a better graphical model can be trained and evaluated. If we have more high-quality data, then we can expect to be able to train a more consistent and accurate capsicum detector. Second, while we can detect capsicums at pixel level in this paper, we can not provide a bounding box or where the centre of the object is. As such, we will develop a blob detector that can potentially distinguish between adjacent crops.

Acknowledgment

The authors would like to thank David Carey, Brian Fisher, Raymond Russell, and Jason Kulk for their invaluable assistance with data collection as well as Andrew English and Alex Bewley for their valuable comments and feedback.

References

[1] The United Nations, "Food production must double by 2050 ." http://www.un.org/press/en/2009/gaef3242.doc.htm, 2009.
[2] High Level Expert Forum, "How to Feed the World in 2050," 2009.
[3] A. English, P. Ross, D. Ball, and P. Corke, "Vision based guidance for robot navigation in agriculture," in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pp. 1693–1698, IEEE, 2014.
[4] C. W. Bac, J. Hemming, and E. J. Van Henten, "Robust pixel-based classification of obstacles for robotic harvesting of sweet-pepper," *Comput. Electron. Agric.*, vol. 96, pp. 148–162, Aug. 2013.
[5] K. Yamamoto, W. Guo, Y. Yoshioka, and S. Ninomiya, "On Plant Detection of Intact Tomato Fruits Using Image Analysis and Machine Learning Methods," *Sensors*, 2014.
[6] S. T. Nuske, S. Achar, T. Bates, S. G. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," in *Proceedings of the 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '11)*, September 2011.
[7] Q. Wang, S. T. Nuske, M. Bergerman, and S. Singh, "Automated crop yield estimation for apple orchards," in *13th Internation Symposium on Experimental Robotics (ISER 2012)*, no. CMU-RI-TR-, July 2012.
[8] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pp. 5314–5320, Nov 2013.
[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1, pp. 886–893 vol. 1, June 2005.
[10] A. Ng, "Sparse autoencoder," *CS294A Lecture notes*, vol. 72, 2011.
[11] G. Zhao, T. Ahonen, J. Matas, and M. Pietikainen, "Rotation-invariant image and video description with local binary pattern features," *Image Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 1465–1477, 2012.
[12] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 7, pp. 971–987, 2002.
[13] J. Domke, "Learning Graphical Model Parameters with Approximate Marginal Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2454–2467, 2013.
[14] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.
[15] C. K.-Y. Hung, *Class-Based Object Detection and Segmentation in Low-Altitude Aerial Images*. PhD thesis, The university of Sydney, 2013.
[16] C. D. Manning, *An Introduction to Information Retrieval*. Cambridge University Press, 2009.
[17] C. McCool, C. Lehnert, D. Hall, B. Upcroft, and T. Perez, "Queensland DAFF Strategic Investment in Farm Robotics (SIFR) Milestone Report - Studies on Weed Identi cation and Weed Destruction," tech. rep., QUT, 2015.